WHITE PAPER

DEVELOPING AN ENTERPRISE-WIDE DIGITAL REPOSITORY FOR PROJECT INFORMATION

Developed at the
NASA Goddard Space Flight Center

by
Gail Hodge (Information International Associates)
& Ed Rogers, Systems Management Office, Code 300

May 17, 2004

SUMMARY OF CONCEPT

During the activities related to an enterprise as large as NASA, a tremendous amount of information will be generated. This information includes the results of scientific activities and information about the unique engineering and project management tasks involved in planning, designing, implementing, launching, and maintaining these exploratory missions. NASA has the opportunity to significantly improve the interoperability and long-term usefulness of the information it generates by providing for better management of the information from the beginning of its life cycle. Such improved management will make valuable knowledge assets of all kinds available now and in the future for the benefit of researchers, project managers, policy makers, educators and the general public.

Historically, project documentation has been managed by each individual project within an Enterprise. The project documents (text, images, video clips, software, etc.) are stored and managed in project libraries using commercial off-the-shelf, internally developed or contractor developed systems. This practice provides the project manager with flexibility within the NASA guidelines and direct access to the documents during the course of the project. However, the diversity of systems results in the compartmentalization of project documentation. Valuable knowledge assets may be lost, especially when the people involved with the project, who acted as information gatekeepers, are no longer available to perform this function. This approach to project management leads to a lack of interoperability, limited access to knowledge assets in the short term, and the inability to use this explicit knowledge in support of a learning organization.

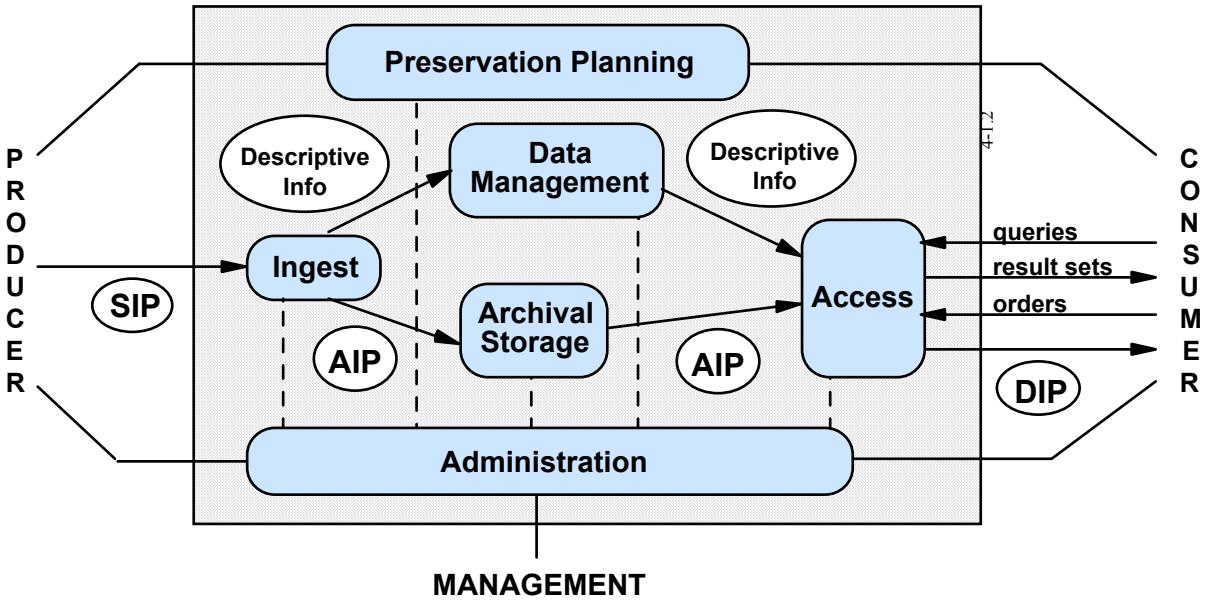
In addition to the significant impact that a digital archiving and preservation system could have on NASA's ability to share information and create a learning organization, such an approach to government information management is called for in the E-Government Act of 2002 Section 207 (E-government Act, 2002). In order to implement the Act, the Office of Management and Budget has created several committees including the Interagency Committee on Government Information (CIO Council, 2004). The major components of a successful government information archive outlined below are also under development by working groups under this committee. These requirements are also the specific embodiments of the high-level requirements of the Federal Enterprise Architecture, particularly the Data Reference Model.

The following series of discussions describes six components of a successful archiving system: 1. Overall architecture based on the Open Archival Information Systems Reference Model (OAIS RM), 2. Ingestion (an agreed upon mechanism for ingesting information into an institutional repository), 3. Archival storage (media stability and preservation), 4. Data management including a metadata structure, taxonomy and persistent identification, 5. Retrieval (search capability & rights management), and 6. Repository management including policy formulation, enforcement and sustainability. The first component, the OAIS RM is an overarching architecture and an ISO Standard that provides a framework to describe the remaining components. Both safety and mission success in NASA require a comprehensive Enterprise wide policy for electronic preservation. This paper lays out the critical considerations necessary to make sure that electronic means are used to support new exploration initiatives. A smart archive system does not replace Configuration Management or Records Management. It supports those tasks making their output valuable to the Enterprise, the Program and the Agency.

E-government Act of 2002. (Public Law 107-347, 44 U.S.C. Ch 36)
http://www.cio.gov/archive/e_gov_act_2002.pdf
CIO Council. Interagency Committee on Government Information.
http://www.cio.gov/documents/ICGI.html

1.0 Open Archival Information System Reference Model: A Framework for Capturing, Storing and Accessing Enterprise-Wide Information

It is important to have an overall strategy for managing and communicating about the repository. The Open Archival Information System Reference Model (OAIS RM) (ISO 1472) provides such a framework by laying out the concepts and defining the terms to communicate about a long-term repository, either digital or analog (CCSDS, January 2002). The Consultative Committee on Space Data Systems (CCSDS) originally developed the OAIS RM for the space data community. However, it was soon acknowledged as a generalized reference model. The basic functional components of the OAIS RM are presented in the figure below.



SIP = Submission Information Package
AIP = Archival Information Package
DIP = Dissemination Information Package

Data or information packages are contributed to the repository either through submission by the producer or harvesting by the repository (Ingest of Submission Information Packages). The content is stored in the repository (Archival Storage of Archival Information Packages) while any accompanying descriptive information is created or extracted and structured to support searching and administration of the objects (Data Management). The Customer executes search queries against the repository (Access). Depending on access controls, the results are repackaged and presented to the customer. Administration involves the day to day operations of the archive and the Preservation Planning activity looks toward the long-term needs, implementing and adjusting preservation strategies as necessary. Management provides

resources and sets the general direction for the repository in order to meet the needs of the Enterprise.

The OAIS RM is the basis for many repositories of electronic resources within the scientific publishing and research library communities. OAIS-compliance is a requirement for the National Archives and Records Administration's Electronic Records Archive development. The OAIS RM is also being used to redesign the National Space Science Data Center (NSSDC) headquartered at GSFC (Sawyer, 2003). Other major archives using the OAIS RM include the NASA Life Science Archive, the National Snow and Ice Data Center, the National Oceanographic Data Center, several archives within the European Space Agency, and the Australian National Archive.

The OAIS RM will be used to frame the remainder this series of papers, with each paper highlighting the components needed in one of the major functional areas in the RM – ingest, data management, archival storage, access, and administration and preservation planning.

CCSDS. (January 2002). "Reference Model for an Open Archival Information System (OAIS). CCSDS-650.0-B-1 Blue Book. http://ssdoo.gsfc.nasa.gov/nost/wwwclassic/documents/pdf/CCSDS-650.0-B-1.pdf

Sawyer, D. (2003). "NSSDC Role and OAIS Implementation." Presentation to the Library of Congress, June 2003. http://ssdoo.gsfc.nasa.gov/nost/isoas/presentations/oais_nssdc_implementation_200306.ppt

2.0 Creating an Enterprise Repository (Ingest)

Central to the development of an enterprise repository is the ability to know about or establish control over knowledge assets through interaction between the producers or creators of the project information and the repository. This is especially critical in a large Enterprise with disparate projects.

Interoperability and long-term preservation of knowledge assets from a new Enterprise can be improved by implementing an Enterprise-wide documentation management system. However, a single solution may not be appropriate for all projects in such a large Enterprise. In addition, over the long term of the Enterprise this approach will not adequately address the anticipated technological changes. A more flexible way to ensure interoperability across project libraries in the Enterprise is to establish a framework for creating an institutional repository in which the Enterprise has some level of control over critical knowledge assets.

Based on a detailed review of the ingest process and the interface between the producer and the repository, the CCSDS produced a general framework for the producer-repository interaction, which acts as a check list for what should be considered when negotiating with the producers on how the repository will be created (CCSDS, December 2002). The CCSDS Methodology suggests that an organization, like the Exploration Enterprise, would tailor the Methodology to create a community-specific implementation, including further definition of terms, the creation of an information model for the community, and identification of particular standards and tools to be used in the repository-building process. The Exploration Enterprise Methodology would provide details on how the components described in the remainder of this white paper series would be implemented.

CCSDS. (December 2002). "Producer-Archive Interface Methodology Abstract Standard. CCSDS-651.0-R-1. http://ssdoo.gsfc.nasa.gov/nost/isoas/CCSDS-651.0-R-1-draft.pdf

3.0 Storing the Objects in the Repository (Archival Storage)

Decisions made when taking objects into the repository will impact the ability to access, preserve and reuse the object in the future. These decisions involve what will be stored in the repository, what formats will be accepted and what media will be used.

Depending on the agreement between the project and the Enterprise (the producer and the archive), the actual digital objects in the repository may be held physically or virtually. The approach may differ depending on the project, the format types, the project phase, or other factors. A hybrid system can be envisioned where some project documentation within the Enterprise is held physically in the repository while other documentation remains with the producer (i.e., the project) or another third-party. In this case, it is important that the Enterprise repository have knowledge of the existence of the object in order to provide proper access and, together with the producer, provide for long-term preservation. Eventually, the content may be brought under the Enterprise repository's control when the project is completed.

Issues related to storage include the stability of the media, the acceptable formats and whether they will be stored in their native format or transformed into another form (for example conversion of a Word document into pdf). Proposed standards such as PDF-Archive from Adobe must be considered. (Hodge & Frangakis, 2004).

Hodge, G. & E. Frangakis.(2004) "Digital Preservation and Permanent Access to Scientific Information: The State of the Practice." CENDI-2004-3:Rev. 05/04. http://www.dtic.mil/cendi/publications/04-3dig_preserv.pdf

4.0 Data Management: Metadata, Terminologies and Persistent Identifiers

Data management provides the necessary information to describe, access and locate objects in the repository for immediate and long-term use.

4.1 Metadata

Metadata (data about data) describes each resource so it can be discovered and evaluated by the user. It also supplies elements that aid in the administration of the repository; and maintains the structure of the repository by providing linkages among related objects. In the case where the digital object is not brought into the repository or where the metadata serves as a surrogate for a non-text object such as an image, the metadata is critical. A central metadata repository containing key information can also facilitate a single search across multiple repositories containing different types of objects that cannot easily be searched otherwise.

The GSFC Library has developed a core set of metadata elements – the Goddard Core -- for the discovery of project-related objects. This element set is based on the Dublin Core Metadata Element Set (ISO 15836) (DCMI, 2003). Elements have been added to accommodate key project information such as project name and instrument or platform name. For long-term preservation, additional metadata elements will be included for version control, checking files for bit loss after transformation of the content to a different format or following migration from one

medium to another, and documenting the significant characteristics and behavior of the object so it can be rendered in a new technology environment in the future. These elements are still being defined by the PREservation and Metadata Implementation Standards Group (PREMIS, 2004).

The Goddard Core is being tested in the GSFC Library's Digital Archiving System and further use across the Center is being discussed with the GSFC Webmasters. Similar efforts are underway at Glenn Research Center, the Jet Propulsion Laboratory, and the Kennedy Space Center. These groups are working together and with others such as the Boeing Company, the UN Food and Agriculture Organization, and the University of Maryland, College of Information Studies, to explore a more universal core set of metadata elements for projects.

Other users within NASA could modify the Goddard Core by developing an Enterprise or project-specific profile or extension. Knowledge management, library, and XML experts within the Enterprise should jointly develop a common metadata approach that can then be expressed as an XML schema.

A key issue is how the metadata will be created. The metadata can be submitted actively by the producer or the producer can provide a specially formatted file containing the metadata that the repository harvests. In the latter case, the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) is emerging as a standard [OAI, 2002]. It is widely used in the creation of institutional and discipline-oriented repositories including the ArchivX e-print server. The producer uses PMH tools to expose OAI-compliant metadata in a way that it can be harvested and merged with an existing central repository.

OAI is the basis for the redesigned NASA Technical Report Server at LARC, the D-Space Institutional Repository developed by Hewlett Packard and MIT, the Digital Library of Earth System Education (DLESE), and the National Science Digital Library at Cornell. (Other implementations are described in Van de Sompel & Lagoze, 2003). The GSFC Library has implemented OAI to provide metadata from its IMAGES Database of scientific images and animations to the NASA Image eXchange.

The OAI has a small set of required metadata elements taken from the Dublin Core. The Goddard Core contains the minimum OAI elements and other OAI elements could be defined by the Exploration Enterprise community.

DCMI (2003). Dublin Core Metadata Element Set. Version 1.1 Reference Edition. http://www.dublincore.org/documents/dces/

OAI. (2002) The Open Archives Initiative for Metadata Harvesting, Version 2.0. http://www.openarchives.org/OAI/openarchivesprotocol.html

Von de Sompel, H. and C. Lagoze. (2003). "Notes from the Interoperability Front: A Progress Report on the Open Archives Initiative." Paper presented at ECDL 2003. Web site available at: http://www.openarchives.org/documents/ecdl-oai.pdf

4.2 Taxonomy, Controlled Terminologies and Other Knowledge Organization Systems

Taxonomies, controlled terminologies and other knowledge organization systems allow the user to consistently find relevant information based on topic or some other important aspect. Knowledge organization systems provide structure for a collection of objects. There are a variety of KOSs that range from simple authority files, such as a list of countries or state names, to schemes that provide extensive information about the relationships between concepts within a discipline or field of activity, such as ontologies, topic maps and semantic networks. [Hodge 2000] A draft taxonomy of KOSs has been developed [Hill & Hodge 2000].

Significant efforts are underway within NASA and government-wide to develop taxonomies that can be used to organize digital information for particular audiences. These are often expressed in the organization of portals.

NASA continues to develop a series of portals to communicate internally and externally with partners, vendors, contractors and the public. In each instance, thought must be given as to how to express the major aspects of the content in ways that will improve navigation and access for the particular group. At the same time, there is a need to repurpose information for dissemination to multiple groups, through multiple portals and over time. This requires effort to bridge or map the taxonomies at the various levels….project, enterprise, agency-wide, government-wide and then across communities of practice, which may span all these political entities and multiple sectors.

The GSFC Library and the KM Office have been working with the NASA-Wide Taxonomy development team, led by the Jet Propulsion Laboratory, to pilot the implementation of the NASA-Wide Taxonomy [Dutra & Busch]. The NASA-Wide Taxonomy has been implemented in the GSFC Library's Digital Archiving System, where one or more terms from the taxonomy are assigned to each video, image, project document, or web site included in the system. In addition, terms from the Earth Observing System taxonomy pilot were mapped to the NASA-Wide Taxonomy to ensure that the local taxonomy could integrate/nest with the NASA-Wide Taxonomy. This integration allows relevant materials from a local system to be repurposed for a portal that uses the NASA-Wide Taxonomy.

Hodge, G.. "Systems of Knowledge Organization for Digital Libraries: Beyond Traditional Authority Files" CLIR Pub91. April 2000. (www.clir.org/pubs/abstract/pub91abst.html)

Hill, L. and G. Hodge. Taxonomy of Knowledge Organization Sources/Systems. NKOS. July 2000. (http://nkos.slis.kent.edu/KOS_taxonomy.htm **)**

Dutra, J. and J. Busch. (2003). See: http://nasataxonomy.jpl.nasa.gov/

4.3 Persistent Identification

Broken links are a major barrier to long-term use of digital information, which requires consistent since users require consistent, reliable, and accurate access. Addressing this problem requires a persistent identifier that tracks an object regardless of its physical location or current ownership.

Currently, most digital objects on the Web are identified using the Uniform Resource Locator (URL). Unfortunately, this approach directly associates the name of the digital object with a physical location. When the object is removed from its original location or the system is

reorganized, the association between the name and the location of the object is "broken" and accessing the original name yields a 404 error message.

A persistent identifier is similar in function to a Social Security Number which is assigned to an individual and does not change when that person's address changes. Likewise, digital object identifiers must be unique, persistent, independent of specific Web domain names, resolvable using standard Web protocols, and flexible enough to allow efficient management of digital information and accommodation of technological changes.

Two primary persistent identifier systems have emerged: the Persistent URL (PURL) and the Handle System®. Both systems are in use in the government and private sectors to enable Web applications to redirect users from the "persistent URL" to the current location of the object. The PURL is used by the Government Printing Office and the Department of Energy's Office of Scientific and Technical Information. Handles support a federated system of resolvers and can resolve to multiple locations or multiple versions of an object. The Handle System has been adopted by major publishers for persistent identification of commercially traded content through the Digital Object Identifier (DOI) system. Handles are also used by the Defense Technical Information Center, the Library of Congress and the National Agricultural Library.

It is not sufficient to create identifiers and leave them without maintenance; active management is needed in order to gain the benefits of such a system. The resolver must be kept up-to-date with the current URLs for the locations of the objects. While some of this updating can be automated, responsibility for this updating and ensuring its reliability must be assigned.

Establishing methods for persistent identification requires extensive analysis of issues such as preferred identifier approaches, core metadata, identifier maintenance, and relationships with existing information management systems. Consideration must be given to all aspects of the government information life cycle from creation to ultimate disposition, including permanent preservation.

A recent white paper by CENDI, an interagency group of information managers from the major scientific and technical mission agencies, highlights the need for persistent identification [CENDI, 2004]. CENDI and others are working with the Office of Management and Budget's Interagency Committee for Government Information (part of the response to the E-government Act of 2002) to identify the requirements for persistent identification in the government information environment.

CENDI Persistent Identification Task Group. (2004). "Persistent Identification: A Key Component of an E-government Infrastructure." http://www.dtic.mil/cendi/publications/04-2persist_id.pdf

5.0 Search and Rights Management (Access)

The Enterprise Repository requires a search engine and rights management/access controls to provide appropriate access to the repository's contents in support of user queries.

5.1 Search

A major technology component of a digital repository is the search engine. The GSFC KM Office recently identified requirements for such a search engine, including a simple search interface that can be tailored for specific communities of practice, integration with portal and content management systems, the ability to search both metadata and full text (where available), and incorporation of a pre-defined taxonomy or other knowledge organization system.

The most common mode for searching is typed text (entering a keyword or phrase in a search box). However, this may not be appropriate for all situations, content types or formats. Additional modes, such as searching images based on visual elements, color, design, etc. or accessing objects via speech, may require the integration of different search engines.

After the user has searched the repository, visualization and other post-processing tools are needed to allow users to better understand, process, and integrate the results of a search with other tools, with their workflow, and in response to a particular need at hand.

An Enterprise-wide solution to searching is enhanced if it can search across multiple heterogeneous systems, formats and content types using a "single search box". The GSFC Library has implemented WebFeat, which searches across 50 heterogeneous databases, both internal and external to the Library, from a single search box. Other libraries, including LARC have implemented or tested similar systems from other vendors.

5.2 Rights Management

An Enterprise Repository must comply with the laws and regulations concerning access to sensitive information. Rights management is the metadata and processes that ensure that an object is provided only to a user who is eligible to view it. Rights management addresses classified, unclassified but limited distribution material such as EAR and ITAR, and copyrighted and other proprietary materials.

In the past, much of the responsibility to ensure rights management and access control was in the hands of security officers, information managers and librarians, supported by markings requirements. However, in the digital environment, these processes must be supported electronically. The method of rights management recommended is one based on roles. This requires individuals to be marked by their roles in the system. The Repository then marks records by role access privileges. This type of system requires up-to-date and accessible personnel information on assignments, security classifications and access authority something the Agency is already working on.

Digital rights management and access control require that the information be properly marked. This can be done internal to the object or in accompanying metadata. There are several schemes being developed for rights management in a digital environment, many of these initiatives focus on the needs of the commercial sector, specifically the recording and entertainment industries. Complementary efforts must be developed within government. A group under the auspices of the Interagency Committee on Government Information is assessing the marking of web sites. Secondly, there must be a system in place to authenticate the user and to match the user's access profile against the object's distribution markings.

6.0 Administration and Preservation Planning: Setting Repository Policies

Sound administration and preservation planning is critical to the development, successful implementation and sustainability of an Exploration Enterprise repository. The administration function is charged with day-to-day management of the repository, while preservation planning takes the long-term view. Both functions, along with management, the producers and the customers, must jointly establish policies that are both efficient and effective.

In the area of Preservation Planning, preservation strategies are critical and must be tailored to the current and long-term uses of the repository, the content types and formats included, and the growing body of appropriate practices in the area of digital preservation and long-term access. (Beagrie & Jones, 2001)

Preservation is key to building an effective learning organization as called for in the CAIB Report. Without timely and convenient access to Enterprise wide knowledge, individuals, teams and projects will not be able to effectively draw upon the lessons learned by those who work around them or across the Agency. Therefore, it is crucial that standards for creating the repository are program/project independent. The Enterprise needs to be the authority on repository management including auditing compliance. Currently, individual projects determine their repository needs. This has created the situation where large amounts of information already generated in project cycles is unavailable for new programs. Code T faces this situation in trying to collect the lessons learned from prior lunar exploration missions.

Ideally, the authority for determining Configuration Management, Records Management and Repository Management should reside with the Enterprise and Center that manages the programs. A joint oversight by Center and Enterprise will help to ensure that funding support is adequate to sustain the preservation efforts, that projects undertake sufficient time to plan ahead for the preservation of their work and that any special or proprietary systems are fully compatible with the Enterprise preservation architecture. These are costs the Enterprise will pay over and over again if not addressed up front in planning.

The Goddard Space Flight Center is moving to establish the resources of the Library and Information Services Branch as the Center Digital Repository for Project Information. At the same time, the authority for establishing preservation standards is being migrated to Center level control. Goddard has brought together the Library, the Knowledge Management Office and the Office of the Chief Information Officer to address these issues at the Center level. Use of this model and the accompanying experience base of addressing a Digital Repository for Project Information helps position Goddard to support NASA in the formation of advanced Enterprise and Agency Wide Digital Repositories.

Beagrie, N. and Jones, M. (2001). Preservation Management of Digital Materials: A Handbook. http://www.dpconline.org/graphics/handbook/index.html